IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13453

Detecting Malware Activity Using Machine Learning

Prathamesh Jadhav[1], Prathamesh Bhavsar[2], Sanket Deore[3], Kiran Kuyate[4], Miss. Gayatri Bendale[5]

Computer Engineering Dept., Matoshri College of Engineering & Research Centre, Nashik, India[1-5]

Abstract: Malware detection is a critical component of modern cybersecurity, as malicious software poses a substantial threat to the security and privacy of individuals and organizations. Traditional signature-based approaches to malware detection have limitations in identifying new, previously unseen malware variants. Machine learning has emerged as a powerful tool in this domain, offering the ability to detect malware based on patterns and behaviour's rather than relying solely on known signatures. These abstract highlights the key aspects of using machine learning for malware detection. Machine learning algorithms are capable of analysing large datasets of file characteristics, network traffic, and system behaviours to identify subtle and evolving patterns associated with malware. By employing techniques such as deep learning, decision trees, and support vector machines, these algorithms can generalize from labelled training data to make predictions about the nature of unknown files or activities. Additionally, feature engineering and feature selection processes enhance the ability of machine learning models to distinguish between benign and malicious entities effectively. The dynamic nature of malware necessitates real-time or near-real-time detection methods. Machine learning enables the development of predictive models that continuously adapt to new threats, making it possible to stay ahead of evolving malware variants. Moreover, the integration of machine learning with other security measures, such as anomaly detection and threat intelligence, further enhances the overall efficacy of cybersecurity systems.

Keywords: Malware detection, Machine learning, Behavioural analysis, Decision trees, Feature engineering.

I. INTRODUCTION

In the rapidly evolving landscape of cybersecurity, the importance of malware detection cannot be overstated as it serves as a critical Défense mechanism against an array of evolving threats that jeopardize both individuals and organizations. Traditional methods that rely on known signatures and patterns have encountered limitations, as they struggle to keep up with the ever-increasing complexity and sophistication of new malware variants. In response to this challenge, the integration of machine learning techniques has emerged as a crucial and effective solution. Machine learning algorithms have the capacity to examine vast datasets, including file properties, network behaviours, and system activities, thereby enabling the identification of malware through patterns and behaviours rather than static signatures. This dynamic approach not only improves the efficiency and precision of malware detection but also provides real-time adaptability, which is vital in combatting the continuous evolution of cyber threats. This chapter delves into the motivation behind developing a robust and adaptable system that can autonomously recognize and address emerging threats, contributing significantly to the protection of digital systems and networks in our interconnected world. By focusing on the problem statement and objectives, the chapter sets the stage for the development of an advanced system designed to safeguard against the ever-changing landscape of malware threats. Such a system plays an essential role in fortifying the cybersecurity of both private and public sectors, ensuring the integrity and security of information and resources in an increasingly digitalized era.

II. PROBLEM STATEMENT

Traditional signature-based approaches have proven inadequate in addressing the everincreasing sophistication and diversity of malware strains. This necessitates the development of a robust and adaptable system that can autonomously identify and combat emerging threats. The central challenge lies in designing a system capable of efficiently analysing and interpreting complex datasets to distinguish between benign and malicious activities, while ensuring real-time adaptability to stay ahead of evolving malware variants.

The system should also seamlessly integrate with existing security measures and threat intelligence sources to offer comprehensive protection against a dynamic and constantly changing cyber threat landscape. The development of such a system represents a critical Endeavor in the ongoing battle to safeguard digital assets and privacy in an increasingly interconnected and vulnerable digital environment.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13453

III. LITERATURE SURVEY

The literature review in this chapter looks into the field of malware identification and analysis utilizing machine learning approaches. It begins by examining the work of B. TAHTACI and B. CANBAY, focusing on the development of machine learning models leveraging n-gram features from decompiled Android packages, crucial for addressing the escalating threat of Android malware. Following this, I. Firdausi, C. Lim, A. Erwin, and A. S. Nugroho's study emphasizes the need for automated behaviour-based malware detection systems, highlighting the efficacy of machine learning classifiers such as k-Nearest Neighbours, Naive Bayes, J48 Decision Tree, Support Vector Machine, and Multilayer Perceptron Neural Network. Additionally, A. Irshad, R. Maurya, M. K. Dutta, R. Burget, and V. Uher explore feature optimization for runtime analysis of Windows-based malware using machine learning approaches, while K. Sethi, R. Kumar, L. Sethi, P. Bera, and P. K. Patra propose a novel machine learning framework for malware detection and classification, showcasing superior accuracy and efficiency compared to traditional methods. These studies collectively underscore the significance of automated malware scanning solutions and dynamic behaviour-based analysis facilitated by machine learning algorithms to combat the evolving landscape of malicious software threats.

IV. SYSTEM ARCHITECTURE

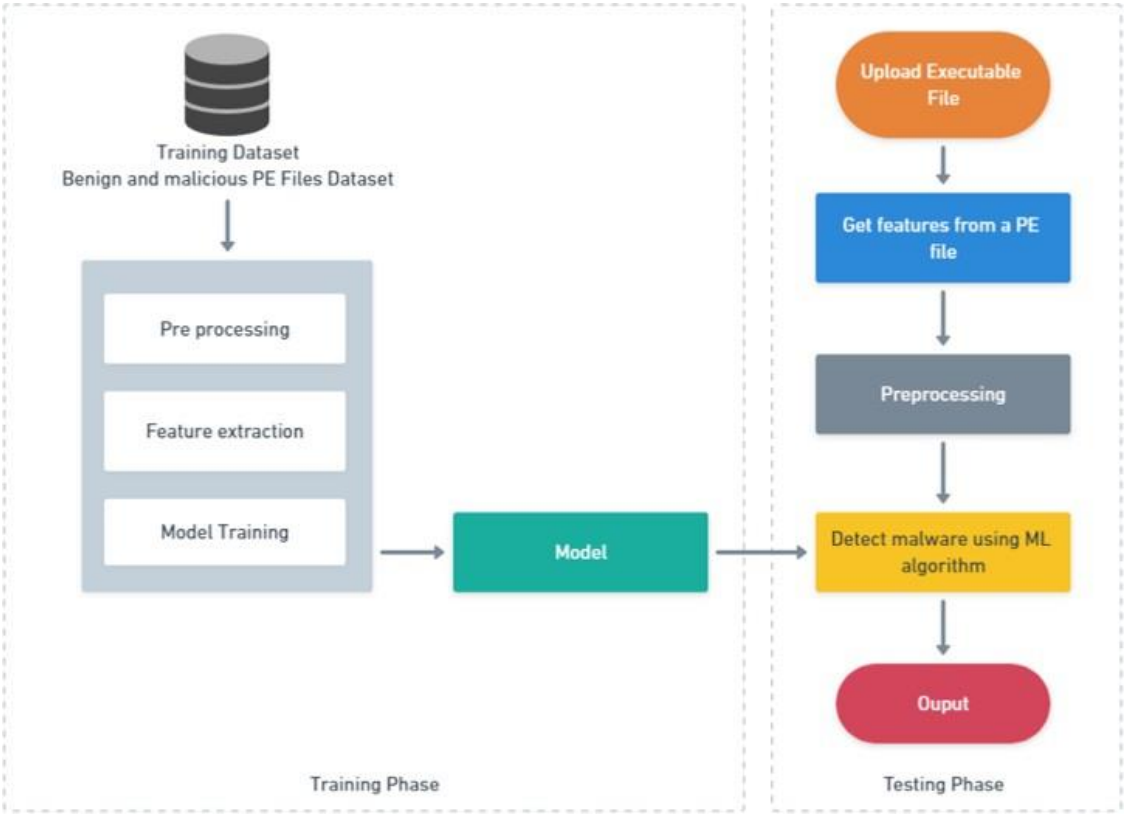

Fig. 1 System Architecture

Detecting malware in Portable Executable (PE) files using machine learning algorithms like decision trees, random forests, and Naive Bayes can be an effective approach. Here's a high-level overview of how you can implement such a system:

- **Dataset Collection**

Gather a diverse dataset of both benign and malicious PE files. You can obtain these files from various sources, including open datasets, malware repositories, and clean software installations.

- **Preprocessing**

Clean and preprocess the data. This may involve dealing with missing values, normalizing features, and encoding categorical data.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13453

- **Feature Extraction**

After the preprocessing process was done, features from the dataset were extracted into a feature vector. We used several word n-gram features such as unigram, bigram, trigram, and combination of unigram, bigram, and trigram. Two-term weighting schemes were used for the feature extraction process. The term weighting schemes used were Bag-of-Words (BOW) and Term Frequency- Inverse Document Frequency (TF-IDF).

- **Classification**

We implemented several machine learning algorithms as the classifier for target classification of hate speech in tweets. Those algorithms are Support Vector Machine (SVM) and Naive Bayes (NB) According to the previous study of hate speech classification, The training phase used 80% of the dataset as the training data, while the testing phase used the remaining 20% of the dataset as the testing data.

- **Evaluation**

The evaluation measurement used in this study is F1- score. Accuracy is not used as evaluation measurement because it cannot guarantee that high accuracy shows that the model can predict well considering the accuracy paradox. F1- score is obtained by calculating harmonic mean between recission and recall.

V. SEQUENCE DIAGRAM

The purpose of interaction diagrams is to visualize the interactive behavior of the system. Visualizing the interaction is a difficult task. Hence, the solution is to use different types of models to capture the different aspects of the interaction.

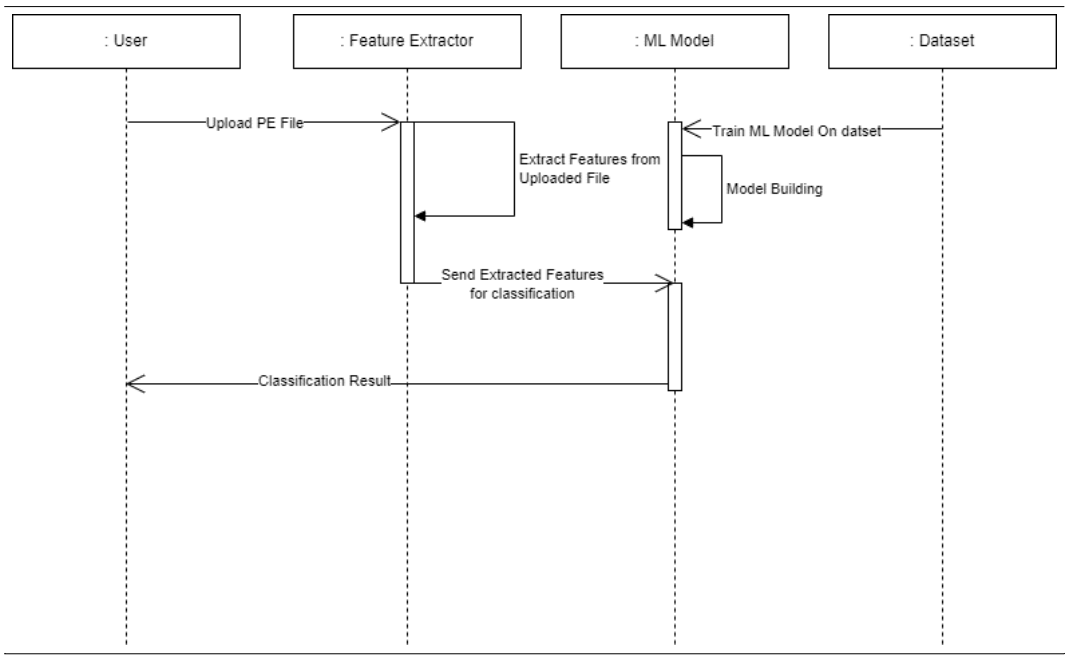

Fig. 2 Sequence Diagram

Purpose of a Sequence Diagram -

- To model high-level interaction among active objects within a system.
- To model interaction among objects inside a collaboration realizing a use case.
- It either models' generic interactions or some certain instances of interaction.

The depicted sequence diagram delineates the primary interactions among the user, the system, and the malware detection engine. In a practical scenario, the user uploads a portable executable file and subsequently extracts the necessary features from the file. The machine learning model utilized for malware detection is typically trained on an extensive dataset comprising known malware samples and benign data, enabling it to generate precise predictions.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13453

VI. PROTOTYPE MODEL OF PROJECT

6.1. HOME PAGE:

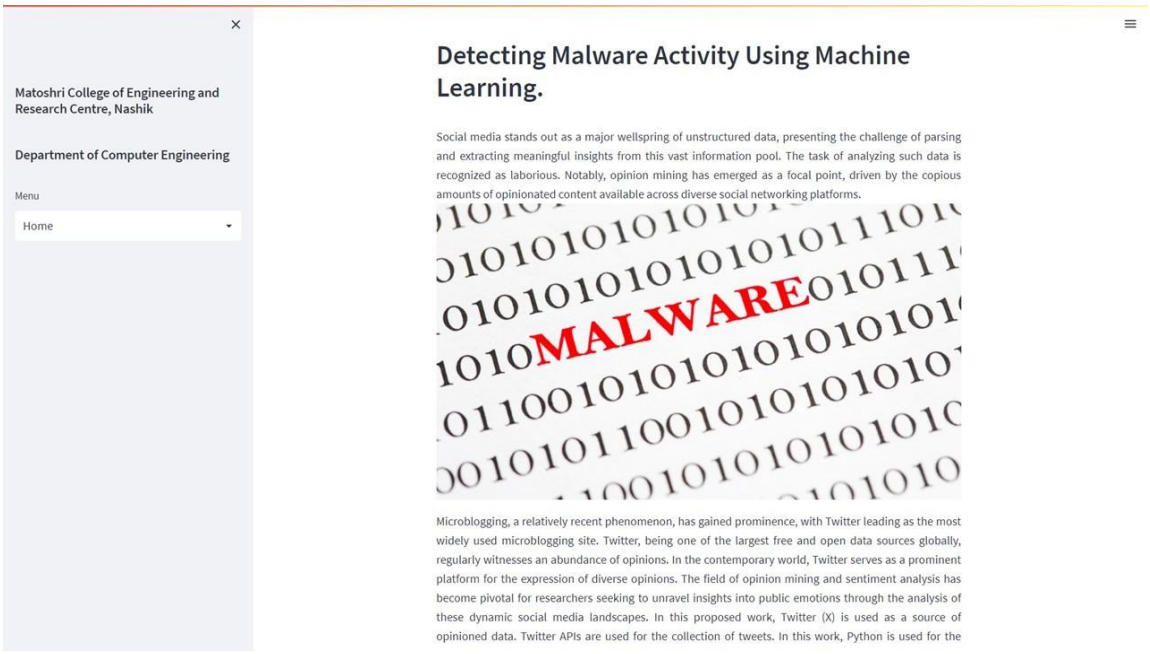

Fig. 6.1. Home Page

6.2. Sign Up Page:

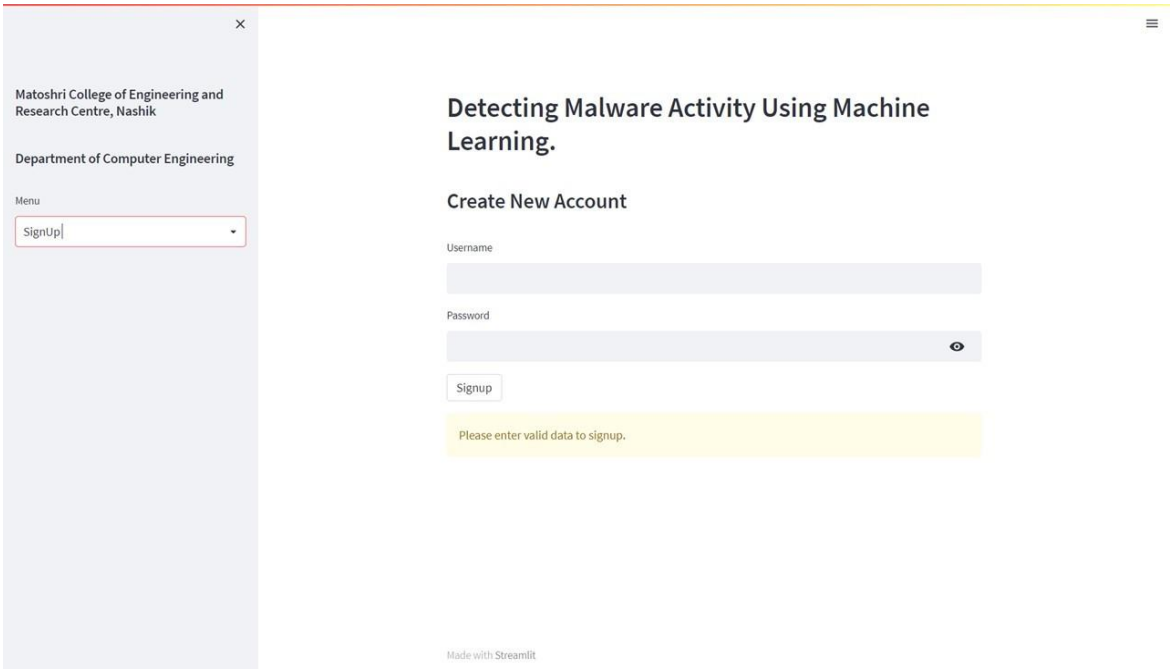

Fig. 6.2. Sign-up Page

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.13453

6.3. LOGIN PAGE:

Fig. 6.3. Login Page

6.4. MALWARE ANALYSIS USING TOTAL VIRUS SET:

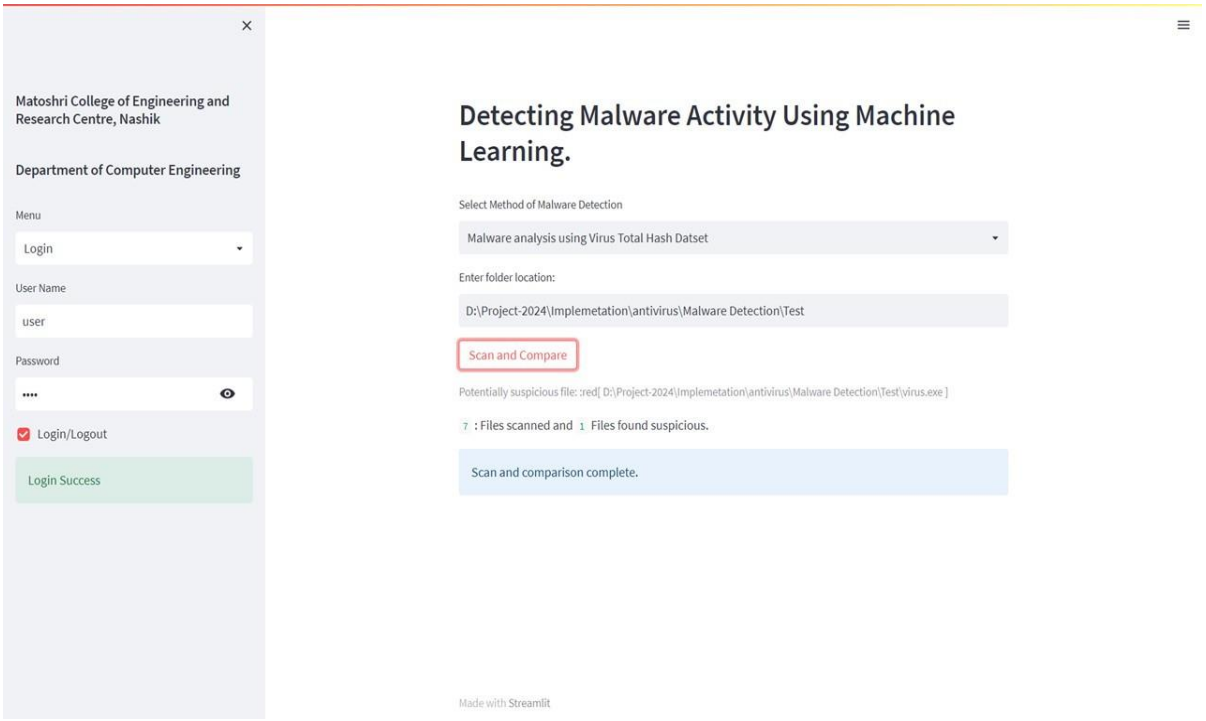

Fig 6.4. Malware Analysis Using Total Virus Set

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.13453

VII. CONCLUSION

The project on detecting malware activity using machine learning stands out with its multifaceted approach to cybersecurity, offering three key features that enhance its effectiveness. Firstly, the system employs machine learning models for malware detection in PE files, leveraging advanced algorithms to identify potentially harmful patterns within executable files. This method improves detection accuracy by focusing on the behavioral and structural aspects of the files. Additionally, the project includes malware detection using the Total Virus hash dataset, enabling the system to quickly identify known malicious files based on hash comparisons and flagging files that have previously been identified as threats, bolstering the system's overall efficiency. The project incorporates a USB drive scanner, allowing for real-time scanning and monitoring of external storage devices. This capability helps prevent the spread of malware through portable media, providing an added layer of protection for users. In sum, these three integrated features create a robust and adaptable system that significantly contributes to safeguarding digital systems and networks against a wide array of cyber threats.

REFERENCES

[1]. B. TAHTACI and B. CANBAY, "Android Malware Detection Using Machine Learning," 2020 Innovations in Intelligent Systems and Applications Conference (ASYU), Istanbul, Turkey, 2020, pp. 1-6, doi: 10.1109/ASYU50717.2020.9259834.

[2]. I. Firdausi, C. lim, A. Erwin and A. S. Nugroho, "Analysis of Machine learning Techniques Used in Behavior-Based Malware Detection," 2010 Second Interna tional Conference on Advances in Computing, Control, and Telecommunication Technologies, Jakarta, Indonesia, 2010, pp. 201-203, doi: 10.1109/ACT.2010.33.

[3]. A. Irshad, R. Maurya, M. K. Dutta, R. Burget and V. Uher, "Feature Optimiza- tion for Run Time Analysis of Malware in Windows Operating System using Machine Learning Approach," 2019 42nd International Conference on Telecom- munications and Signal Processing (TSP), Budapest, Hungary, 2019, pp. 255- 260.

[4]. K. Sethi, R. Kumar, L. Sethi, P. Bera and P. K. Patra, "A Novel Machine Learn- ing Based Malware Detection and Classification Framework," 2019 International Conference on Cyber Security and Protection of Digital Services (Cyber Secu- rity), Oxford, UK, 2019, pp. 1-4, doi: 10.1109/CyberSecPODS. 2019. 8885196.

[5]. B. Alsulami, A. Srinivasan, H. Dong and S. Mancoridis, "Lightweight behavioral malware detection for windows platforms," 2017 12th International Conference on Malicious and Unwanted Software (MALWARE), Fajardo, PR, USA, 2017, pp. 75-81, doi: 10.1109/MALWARE.2017.8323959.

[6]. M. Barat, D. B. Prelipcean and D. T. Gavrilut, "An Automatic Updating Perceptron-Based System for Malware Detection," 2013 15th International Sym- posium on Symbolic and Numeric Algorithms for Scientific Computing, Timisoara, Romania, 2013, pp. 303-307, doi: 10.1109/SYNASC.2013.47.

[7]. S. A. Roseline, A. D. Sasisri, S. Geetha and C. Balasubramanian, "Towards Efficient Malware Detection and Classification using Multilayered Random For- est Ensemble Technique," 2019 International Carnahan Conference on Security Technology (ICCST), Chennai, India, 2019, pp. 1-6, doi: 10.1109/CCST.2019.8888406.

[8]. E. Gandotra, D. Bansal and S. Sofat, "Zero-day malware detection," 2016 Sixth International Symposium on Embedded Computing and System Design (ISED), Patna, India, 2016, pp. 171-175, doi: 10.1109/ISED.2016.7977076.

[9]. S. Gu¨lmez and I. Sogukpinar, "Graph-Based Malware Detection Using Opcode Sequences," 2021 9th International Symposium on Digital Forensics and Security (ISDFS), Elazig, Turkey, 2021, pp. 1-5, doi: 10.1109/ISDFS52919.2021.9486386.